IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 11-15
www.iosrjournals.org

WEB PERSONALIZATION BASED ON ROCK ALGORITHM

Ms. Bhagyashree Ambulkar[1], Ms. Rajeshree Ambulkar[2], Mr. Praful Barekar[3]
[1]*bhagyashree_ghriit@yahoo.com, MCA Dept., GHRIIT. Nagpur, Maharashtra*
[2]*rajeshree28ambulkar@gmail.com, CSE Dept., PIET, Nagpur, Maharashtra*
[3]*praful.barekar20@gmail.com, CT Dept., YCCE, Nagpur, Maharashtra*

Abstract: *With the demand of different information by different users from same web page becomes critical problem for web personalization. The motive of clustering analysis is to discover rich quality of clusters such that the similarity between the intra-cluster is high. By usingpersonalization,access to the web pages or the contents of a Web page are modified to better fit the needs of the user.This may involve actually creating Web pages that are unique per user or using the desires of the user to determine what Web documents to retrieve. The proposed approach aims to mine and improve the search result as per user's need. This paper proposes the novel method for web personalization based on ROCK algorithm for categorical attributes.*
Keywords -*web usage mining, web personalization, hierarchical agglomerative algorithm.*

I. INTRODUCTION

The World Wide Web is the huge collection of information resources on the internet. It is a huge warehouse of many interlinks documents. When user navigates through Internet, he/she gets lots of irrelevant documents after navigating several links. Therefore, the requirement for expecting user needs in order to improve the usability and user custody of a Web site can be addressed by personalizing it.Personalization can be done through clustering, classification or prediction. Through classification, the requirements of a user are determined based on those for the class. With clustering, the requirements are determined based on those usersto which he or she is determined to be similar. Finally, prediction is used to predict what exact requirement of the user.The study demonstrates that our approach is general and effective for mining the web data from categorical attributes for web personalization. In this paper, we study clustering algorithms for both Boolean data and categorical data. We show that traditional clustering algorithms do not always work with boolean and categorical attributes. Clustering analysis finds clusters of data objects that are similar in some sense to one another. The traditional clustering algorithmsuse distances between points for clustering are not appropriate for boolean and categoricalattributes. ROCK (RObust Clustering using linKs) is a clustering algorithm for data with categorical and Boolean attributes[1]. It redefines the distances between points to be the number of shared neighbors whose strength is greater than a given threshold and then uses a hierarchical clustering scheme to cluster the data.[15]The ROCK not only generates better quality clusters than traditional algorithms, butit also exhibits good scalability properties[1].

II. PERSONALIZATION

Personalization is a process of collecting and storing information about website visitors, analyzing the information, and, based on the analysis, supplying the right information to each visitor at the right time[2].

2.1 Elements involved in personalization

- Users- Users have profiles, both individual and group. These profiles contain characteristics (such as location, interests, job description) which can be used to personalize the content they can see. Users also take actions. These can then be analyzed and matched against behavior rules to tailor the content they see.

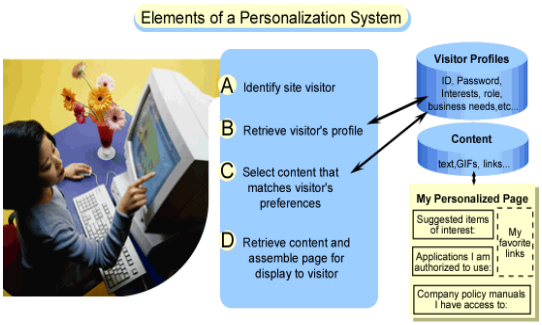

Fig 1: Elements of a personalization system

- Content-Content is what the user wants to see. Contents can be categorized, and therefore made available to users according to predefined rules. must be dynamic; in other words the content must, in some way, be dependent upon the user – if every user would see the same content, then personalization would be redundant[3].
- Rules-define how personalization actually happens and also which content the user can see, and when [3].

2.2 The phases involved in personalization

- Data Preparation: - The worthiness of patterns discovered in the web usage mining process highly depends upon the quality of data used in the mining process[20]. The web page information is stored in the server log file when the web browser traces the web page.Any user browser hits a URL in a domain; the information related to that operation is recorded in an access log file.[21].
- Data Preprocessing: - Data cleaning is also a customized step, which includes integrating different usage logs and parsing data from these usage logs. Filtering is the most important task in web usage mining, since the quality of mined patterns depends upon this directly[2, 4].
- Pattern Discovery: - Pattern discovery is the main issue with both web usage mining & data mining. The search space increases exponentially as the lengths of patterns to be discovered increases [2, 4]. The nature of data to be clustered plays a key role when choosing the right algorithm for clustering.
- Pattern Analysis: - This involves the analysis of navigation patterns and other valuable information extracted from web usage mining before applicable to real world problem.
- Pattern Application: - The last step in the web usage mining is the application of result to convey Recommendations to the user. The discovered knowledge can be used for various applications such as website improvements, web caching, web personalization and business intelligence.

2.3 Types of Personalization

Personalization can be either Explicit-Customization: whereby the user makes selections from a choice of content sources.

- Implicit-Rules based: business managers define specific rules for actions based upon specific profiles and/or behavior.
- Simple filtering - elections are made on the basis of predefined profiles at user and/or group level.
- Collaborative / recommendation filtering: user behavior is registered according to predefined rules. These rules are based on behavior observed with like-minded individuals. The information collected is used to tailor the information displayed to the user, particularly in the form of recommendations [3].

Personalization can be used in many cases, for example in an internet pages, intranet pages, sales and distribution websites, such as Amazon, can combine a user profile, the user's sales history and their browsing history to make suggestions as to what might interest the user next. As well as most of the major search engine websites have very powerful analytical tools which record user behavior, the term which they are search and the websites they really visit. This is then used to customize the content provided particularly with regard to displaying advertisements.

III. RELATED WORK

With the growth of web user and necessity to provide required information to the users efficiently and quickly to the web users becomes very critical. Many web personalization algorithms are recently have been proposed.SeqPAM for clustering sequential data for web personalization [9].Neuro-fuzzy clustering algorithms are used to mine the web server logs for a given period of time using unsupervised and competitive learning algorithm like Kohonen'sself-organizing maps (SOM) and interpreting those results using Unified distance Matrix (U-matrix) [10]. The WordNet-enabled W-kmeans algorithm, an enhancement of standard k-means algorithm which uses the external knowledge from WordNethypernyms and that has been previously used for document clustering, to user profile clustering by analyzing the users' historical data[11]. The K-modes algorithm finds centroids that are valid patterns, truly representative of a cluster, even with nonconvex clusters, and appears robust to outliers and misspecification of the scale and number of clusters [12].

Also many clustering algorithms for large databaseshave been proposed. Among them, BIRCH[7] and ScaleKM[6] compress the large database into subclustes. CLIQUE[8] proposed to find the cluster embedded in subspaces of high dimensional data. PSO usually used to solve continuous optimization problems, but the categorical data are non-continuous. The novel POS k-Modes algorithm overcomes this problem [5].

IV. CLUSTERING TERMINOLOGY

In this section we discuss some terminology used in ROCK clustering algorithm that is based on the notion of neighbors and links.

- Neighbors - A neighbor of a certain object is such an object to which similarity with investigated object is equal to or greater than a predefined threshold [1]. Consider sim(xi,xj) be a similarity function which finds the closeness between the two points xi and xj. Here we assume the vaues for sim is between 0 and 1. The given threshold θ between 0 and 1, a pair of points xi, xj are defined to be neighbors if the following hold :

  $$sim(xi,xj) \geq \theta \qquad (1)$$

- Categorical Data - Categorical variables represent types of data which may be divided into groups. Examples of categorical variables are field of study,nationality, race, sex, age group, and educational level. While the latter two variables may also be considered in a numerical manner by using exact values for age and highest grade completed, it is often more helpful to categorize such variables into a relatively small number of groups.Categorical data typically is of fixed dimension.[22]

- Links- The number of links between two items is defined as the number of common neighbors they have. The algorithms are a hierarchicalagglomerative algorithm using the number of links as the similarity measures rather than a measure based on distance. Traditional Euclidean distance measures are not appropriate for such data and instead, ROCK uses the Jaccard coefficient to measure similarity.[23] This rules out clustering approaches such as K-means or Centroid based hierarchical clustering. While the Jaccard coefficient provides a reasonable measure of the distance between points, clusters are sometimes not well separated and so a new measure of similarity between points was introduced that reflects the neighborhood of a point. If sim(xi, xj) is the similarity between points, xi and xj, and $0 \leq \theta \leq 1$ is a parameter, then

  $$link(xi, xj) = |\{p : sim(xi, y) \geq \theta\} \cap \{q : sim(xj, y) \geq \theta\} \qquad (2)$$

In words, link(xi, xj) is the number of shared neighbors of xi and xj. The idea is that two points will be "close" only if they share a comparatively large number of neighbors. Such a strategy is proposed to handle the problem of "border" points, which are close to each other, but fit to different clusters.

V. THE ROCK CLUSTERING ALGORITHM

ROCK (RObust Clustering using linKs) [GRS99] is a clustering algorithm fordata with categorical and boolean attributes. It redefines the distances between points tobe the number of shared neighbors whose strength is greater than a given threshold andthen uses a hierarchical clustering scheme to cluster the data.[15]The objective of the clustering algorithm is to group together points that have more links.

The ROCK algorithm is divided into three general parts:
1. Obtaining a random sample of the data.
2. Performing clustering on the data using the link agglomerative approach. A goodness measure is used to determine which pair of points is merged at each step.
3. Using these clusters the remaining data on disk are assigned to them.

For a pair of clusters Ki; Kj, let link(Ki ; Kj) store the number of cross links between clusters Ki and Kj, that is,
$\sum p_q \in K_i; p_r \in K_j \; link(p_q; p_r)$. Then, we define the goodness measure g(Ki; Kj) for merging clusters Ki; Kj as follows. The goodness measure used to merge clusters is

$$g(Ki, Kj) = \frac{link(Ki, Kj)}{(ni+nj)^{1+2f(\Theta)} - ni^{1+2f(\Theta)} - nj^{1+2f(\Theta)}} \qquad (3)$$

Here link(Ki, Kj) is the number of links between the two clusters. Also, ni and nj are the number of points in each cluster. The denominator is used to normalize the number of links because large number of cluster would be expected to have more links. $ni^{1+2f(\Theta)}$ is an estimate for the number of links between pair of points in Ki

when the threshold used for the similarity measure is Θ. The function f(Θ) depends on the data, but it is found to satisfy the property that each item in Ki has approximately . $ni^{f(\Theta)}$ neighbors in the cluster.

The first step in the ROCK algorithm converts the adjacency matrix into a Boolean matrix where an entry is 1 if the two corresponding points are neighbors. If adjacency matrix has the size n^2, then there is an $O(n^2)$ steps. The next step converts this into a matrix indicating the links. This can be found by calculating *S X S*, which can be done in $O(n^{2.37})$[GRS99]. The hierarchical clustering portion of the algorithm then starts by placing each point in the sample in a separate cluster. It then successively merges clusters until k clusters are found. To facilitate this processing, both local and global heaps are used. A local heap, q, is created to represent each cluster[13]. Here q contains every cluster that has a nonzero links to the cluster that corresponds to this cluster. Initially, a cluster is created for each point, t_i. The heap for t_i, $q[t_i]$, contains every cluster that has a non zero link to {ti}. The global heap contains information about each cluster. All information in the heap is ordered based on the goodness measure, which is shown in above equation [14].

VI. CONCLUSION

In this paper we discussed the web personalization using ROCK algorithm and tried to explain how the categorical data is clustered by ROCK. It redefines the distances between points to be the number of shared neighbors whose strength is greater than a given threshold and then uses a hierarchical clustering method to cluster the data.[15]At present, personalization is largely limited to closed worlds. The main aim of the web personalization is to provide the user centric information without expecting them ask for it explicitly. Various e-commerce sites that collect data about user preferences have little incentive to make this data available for use by competing services. In future we will do the practical implement of Rock algorithm over the real life example of web search and find the optimize solution for the user.

REFERENCES

[1] Sudipto Guha , Rajeev Rastogi , Kyuseok Shim "ROCK: A Robust Clustering Algorithm for Categorical Attributes".
[2] Amit Rustagi "A Near Real-Time Personalization for eCommerce".
[3] http://dev.day.com/docs/en/cq/5-5/administering/ personalization.html
[4] Dimitrios Pierrakos, Georgios paliouras, Christos Papatheodorou and Constantine D. Spyropoulos "Web Usage Mining as a Tool for Personalization: A Survey" User Modeling and User-Adapted Interaction 13: 311-372,2 003.
[5] Lu Mei, Zhao Xiang-Jun "A Novel PSO k-Modes Algorithm for Clustering Categorical Data".
[6] P. S. Bradley, Usama Fayyad, and Cory Reina "Scaling Clustering Algorithms to Large Databases (Extended Abstract)".
[7] Tian Zhang, Raghu Ramkrishnan, Miron Livny "BIRCH - An efficient data clustering model for very large databases".
[8] Raghunath Kar & Susant Kumar Dash "A Study On High Dimensional Clustering By Using Clique".
[9] Pradeep Kumar, Raju S. Bapi, P. Radha Krishna, SeqPAM: A Sequence Clustering Algorithm for Web Personalization", International Journal of Data Warehousing and Mining (IJDWM).
[10] Menon, K. ; Smart Eng. Syst. Lab., Missouri Univ., Rolla, MO, USA ; Dagli, Cihan H. , "Web personalization using neuro-fuzzy clustering algorithms ", Fuzzy Information Processing Society, 2003. NAFIPS 2003. 22nd International Conference of the North American, ISBN No. 0-7803-7918-7. \
[11] Christos Bouras, Vassilis Tsogkas" Clustering user preferences using W-kmeans".
[12] Miguel Á. Carreira-Perpiñán, Weiran Wang, "The K-modes algorithm for clustering"
[13] "Data Mining: Introductory And Advanced Topics" By Margaret H Dunham
[14] Margaret H. Dunham "Data Mining- Introduction and Advance Topics".
[15] Dimitrios Pierrakos, Georgios paliouras, Christos Papatheodorou and Constantine D. Spyropoulos "Web Usage Mining as a Tool for Personalization: A Survey" User Modeling and User-Adapted Interaction 13: 311-372,2 003.
[16] Neha Saxena "Improving Web Recommendations Using Web Usage Mining and Web Semantics", San Jose State University
[17] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, 1(2):12–23, 2000
[18] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan SIGKDD Explorations, 1(2):12–23, 2000 Web usage mining: Discovery and applications of usage patterns from web data.
[19] Bettina Berendt, Andreas Hotho, and Gerd Stumme Towards Semantic Web Mining.
[20] Anil Sharma1, Suresh Kumar, Manjeet Singh "Semantic Web Mining For Intelligent Web Personalization"
[21] Mali Bayir , A new relative method for mining web usage data, 2006
[22] http://www.stat.yale.edu/Courses/1997-98/101/catdat.htm
[23] Ashwina Tyagi et al. / International Journal on Computer Science and Engineering (IJCSE),"Implementation Of ROCK Clustering Algorithm For The Optimization Of Query Searching Time"
[24] Mandani Kashmira , , Prof.Hemani Shah , "Analysis Of Hierarchical Clustering Algorithm To Handle Large Dataset", International Journal of Advance Engineering and Research Development (IJAERD)Volume 1,Issue 11, November -2014, e-ISSN: 2348 - 4470 , print-ISSN:2348-6406